Research and Development on the BioEncyclopedia

February 2005

Terence Critchlow (ASC PI LLNL)

Tom Slezak (PL)

UCRL-PRES-209902

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

The Biodefense Knowledge Center

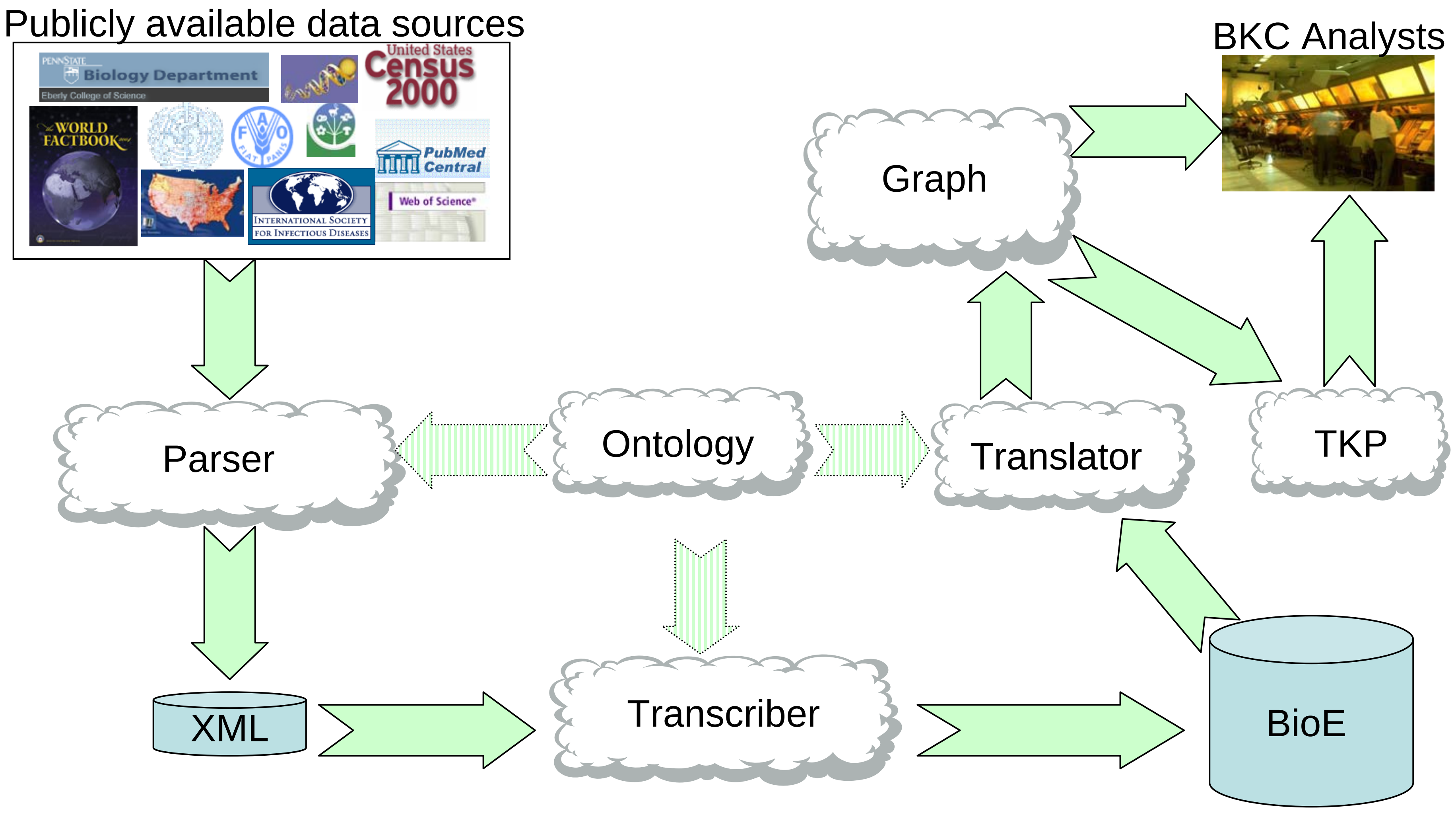
- The BKC Mission is to “Enable collaboration and data sharing among policy makers, responders, analysts, law enforcement personnel, and scientists in the Homeland Security community to assure that timely and authoritative biodefense information is available to those with a need to know.” ***– Bill Colston, BKC Director, August 2004***

The BioE is one of three thrust areas within the BKC technology section

- **Collaborative Visualization**
  - Extending existing tools such as StarLight and InSpire to work within the ADVISE architecture
- **Knowledge management**
  - Applying the Nebraska semantic graph architecture to the BKC domain
- **BioEncyclopedia (BioE)**
  - Integrate all biological information relevant to the BKC mission
  - Develop tailored knowledge products that provide customized views of data
  - Act as a centralized DHS resource for biological information

We are using ASC funding to provide general solutions to research level problems being encountered by the BioE. Program funding is then used to apply those solutions within the BKC environment.

The BioE: Where we are now



Publicly available data sources

PENNSTATE
Biology Department
Eberly College of Science
United States Census 2000
the WORLD FACTBOOK
FAO FIAT PANIS
PubMed Central
Web of Science
International Society for Infectious Diseases

BKC Analysts

Graph

Parser

Ontology

Translator

TKP

XML

Transcriber

BioE

Current status

Ontology Creation

- **Globally consistent view of the world**
- **Created using a graphical interface program developed by the Institute for Human and Machine Cognition**
  - Generates both Java classes and XML
- **Ontology is stored in BioE for use by other programs**

Data Source Parsing

- **Data comes from autonomous publicly available sources**
- **Parsers are manually written to extract data from each source format and convert it to an XML representation**
  - Where needed, the data is transformed to be semantically consistent with the ontology
    - Converting units (inches to cm)
    - Adding implicit information
- **Currently parsing structured and unstructured data sources**
  - RDBMS, HTML, XML, excel, CSV
  - Including (basic) free text data from
    - PubMed
    - ProMed
    - Web Of Science

Current status

Data Ingest

- **The transcriber reads the parser generated XML and uses the ontology to enter it into the BioE appropriately**
  - XML includes provenance information which is also transferred into the BioE
  - Performs concept based canonicalization of data (using simple synonyms)
- **Currently, the BioE is conceptually contained within two databases**
  - Provenance information about where data came from
  - A hyper-normalized view of the data to simplify transfer of information into the semantic graph

Graph Loading

- **The translator is able to sent both the ontology and the associated data into the semantic graph**
  - Uses existing security interface to graph
  - Resolves data type conflicts

Current status

Information Analysis

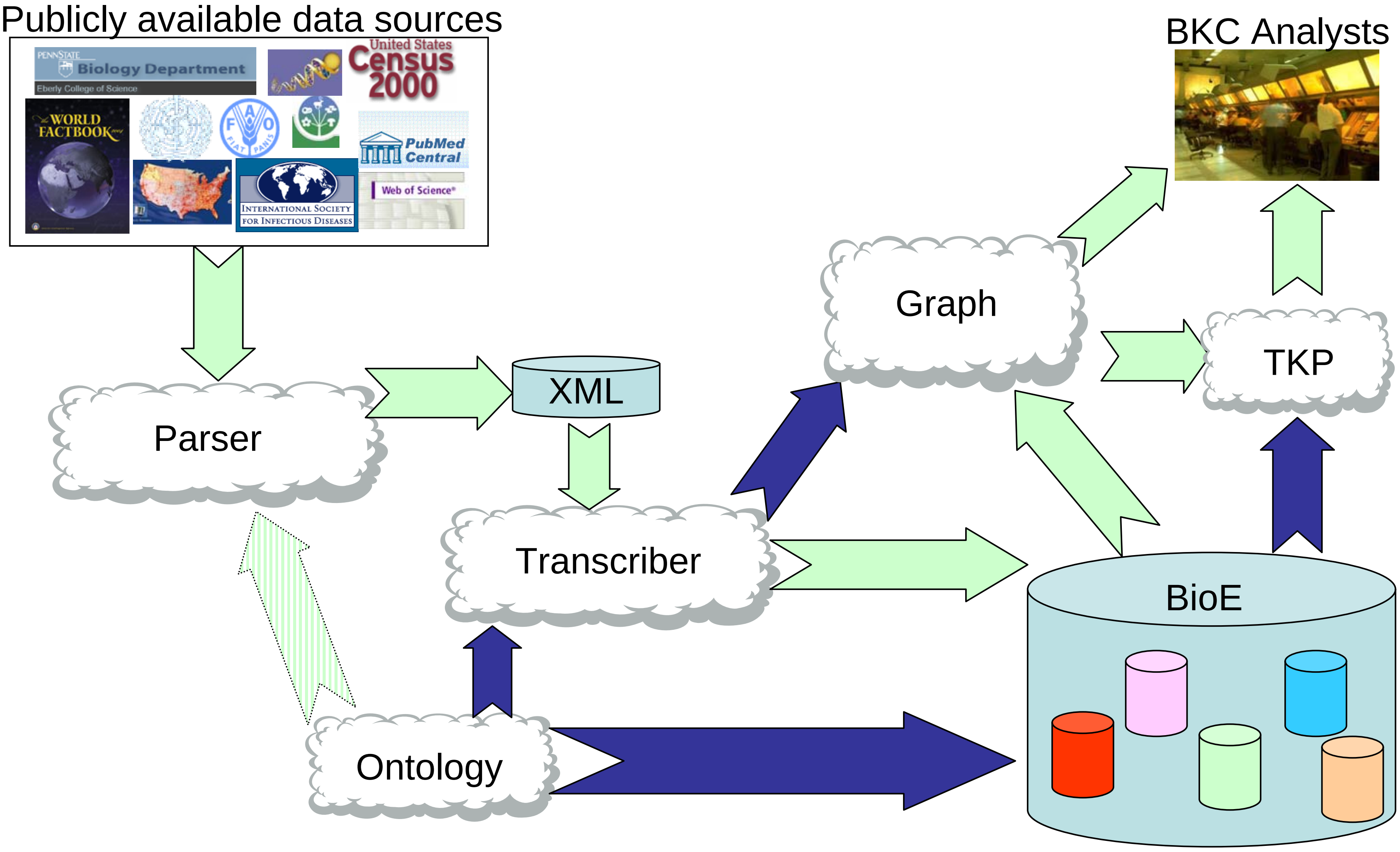
- **Analysts can perform queries directly against the graph through a graph viewer or can perform "*canned queries*" using a pre-defined tailored knowledge product (TKP)**
  - TKPs use a combination of graph queries, detailed information that was not passed to the graph, and data post-processing to provide a straight-forward way to ask a complex question
  - Currently working on two TKPs driven by user demands,
  - Expect to develop several more in the future

Research challenges facing the BioE

- **Need to develop an infrastructure that can support**
  - Complex ontologies
  - Integration of information from a large number of dynamic, heterogeneous, multi-modal data sources
  - Automatic identification and reloading of updated data
  - Incremental changes to the ontology
  - Incremental reloads of data from specific sources
  - Both a relational and graph view of the data

  - Use of external tools to perform data analysis
  - Complex queries through interfaces that are useable by our analysts

The BioE: Where we are going



Publicly available data sources

BKC Analysts

Parser

XML

Transcriber

Graph

TKP

BioE

Ontology

FY05 milestones

- **Change detection algorithms**
  - Prototype, demonstrated on a site of interest to the BioE, that is able to identify when a set of web pages have changed in a significant way and automatically submit updated pages to the appropriate parser
- **Meta-data based tools**
  - Identification of bottlenecks in the BioE data source ingest process
  - Initial demonstration of prototype to resolve one of these bottlenecks (in conjunction with BioE team)
- **Free text analysis and information extraction**
  - Working with other national laboratories and academia
  - Create semantic dictionaries relevant to biodefense domain

Longer term goals

- Continue work on information extraction from free text based on initial identification of problem areas

- Pursue research on automatic identification of relevant data sources

- Ensure scalability of overall architecture

Longer term goals

- **Working as part of BioE team to incorporate:**
  - Constraint-based validation of data (e.g. enumerated types, ranges)
  - Data curation tools
  - Complete data provenance tracking
  - Workflows into TKP framework
  - Multi-modal data
    - Images
    - Chemical structures
  - Complex query evaluation capabilities
    - Document clustering capabilities
    - Sequence similarity algorithms
    - Image analysis techniques

Summary

- There are significant data management needs being faced by DHS programs

- We are developing general solutions to a class of data integration problems

- We have aligned ourselves closely with the BKC effort to ensure our work has immediate impact on DHS

- There is a lot of work left to do

Research and Development on the BioEncyclopedia

February, 2005

Terence Critchlow (ASC PI LLNL)

Tom Slezak (PL)

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.